The Chemical-Biological Coordination Center: An Experiment in Documentation

Raimon L. Beard and Karl F. Heumann

The Connecticut Agricultural Experiment Station, New Haven, and
The Chemical-Biological Coordination Center, National Research Council, Washington, D. C.

IT IS BECOMING A TRUISM among scientists to say that the productivity of past and current research is resulting in an accumulation of data that is difficult to utilize effectively and efficiently. Only the rare scholar can explore a given subject without overlooking some pertinent information; especially is this true for fields at all removed from his own. It is difficult for a cellular physiologist, for example, to find all the units of thought on a given subject scattered throughout the fields of biophysics, biochemistry, and the several subdivisions of biological science. Some data do not reach the usual channels of publication; some are difficult of access because of the multiplicity of journals in all languages; and documentation is particularly incomplete. However excellent the abstracting services may be, they generally do little more than summarize the sense of the papers, and indexing is based on the titles or abstracts, not on the original contribution. Complete indexing of original papers seems beyond the scope of the abstracting services. Indexes alone may not be sufficiently selective for greatest efficiency in searching for information on a restricted topic. Such shortcomings to full documentation have recently been the subject of comment in SCIENCE (**114**, 134 [1951]; **115**, 250; **116**, 19 [1952]) and THE SCIENTIFIC MONTHLY (**75**, 46 [1952]).

The antithesis to full documentation is the attitude that experimentation is easier than a thorough search of the literature, and that it is less expensive to seek answers in the laboratory than in the library. It is startling to find this attitude in high scientific places, for, if carried to an extreme conclusion, it denies the values of scientific communication and ignores our scientific heritage. It is an attitude nourished by the imminent prospect of the scientific literature getting completely out of hand. It indicates an unscientific reluctance to explore possible solutions to the problems of documentation and economy of scientific effort.

One positive approach to facilitating the use of existing information is to be found in the Chemical-Biological Coordination Center of the National Research Council. This organization is engaged in experimentation on the practical problem of assembling and making available information on chemicals and their effects on biological systems. Although in a sense an abstracting service, it is unique in its ideology and in its methods.

The center was set up to serve as a repository for information. By the use of this repository the center expects to encourage and facilitate research activities directed toward the understanding of the interrelationships between chemical structure and biological activity, and between one type of biological activity and another. Underlying the activity is the premise that a biological response to an applied chemical is of interest in its own right, irrespective of the context in which it is found. This follows a growing trend in documentation toward considering units of thought as fundamental, rather than the monograph or the scientific paper that contains the units. Thus the center focuses on two units of thought: One is the chemical unit, which includes the molecular formula and the structural details of a given chemical. The other is the biological unit, which includes certain desired qualitative and quantitative aspects of the biological response induced by an applied chemical. As a more specific (and simplified) example, three "units of thought" are contained in the information that α-naphthylthiourea (ANTU) causes pulmonary edema in Norway rats and, when applied by stomach tube, kills 50 per cent of the experimental animals at a calculated dose of 6.9 mg/kg body weight (LD_{50}). One "unit of thought" includes the molecular formula and structural details of α-naphthylthiourea; one unit includes the information that this chemical induces pulmonary edema in the Norway rat; the third unit indicates the LD_{50} of ANTU for this animal. Intent of the experimenter in observing the response is not overlooked, but it may be considered additional information, supplemental to the observed response. If the ANTU was being studied as a rodenticide, this fact has practical significance and would be noted.

Just as bricks may be used either to pave a terrace or to build a cathedral, items of data can have wider application than to prove the point of an author. It often escapes the attention of an investigator, who gathers and presents his data only in such form as will support his thesis, that they can be used for an entirely different purpose. The breakdown of a paper into units of thought helps to compensate for this oversight, as it makes possible the rearrangement of units to serve any desired purpose.

The desired results, which cannot be achieved without rapid and accurate mechanical aids, dictated the methodology adopted by the center, and to this end a punched-card system was chosen. The application of such a system for the needs of the center required the development of two classifications to permit coding the pertinent chemical and biological information. The end result is, in effect, a file of two sets of punched

cards, which can be searched mechanically for variables. Although search can be made on a single criterion, it is in the search for combinations of ideas, requiring multiple criteria, where this method has a major advantage over conventional indexes. Thus, compounds possessing a specified structure which, in stated amounts, induce a given response in a particular organ system of a given group of animals may be selected from all others not meeting the several requirements.

Editorial comment is often made on the wealth of information that lies dormant in the literature—that Mendel's laws of heredity were long overlooked; that DDT had been synthesized 40 years before it was exploited as an insecticide; that the sulfa drugs were reported long before their general use. No magical system will mine these treasures from the scientific record. The role of the imaginative, inquiring scientist can be facilitated, but not replaced, by these mechanical tools. The punched cards of the center cannot flash out a cancer cure if a cancer cure is not represented by a punched card. They cannot call attention to a new insecticide if the chemicals on file have not been tested on insects. They cannot (now) print a new book by the press of a button. What they *can* do is to make possible the rapid assembly of information on a multiplicity of ideas in combination and to a degree of selectivity not matched by any other existing method. They *can* serve as a research tool in testing hypotheses or seeking generalities from a limited series of observations. They *can* expedite the study of correlations between chemical structure and biological response, or correlations between one type of biological response and other biological events. Such studies can lead to prediction, but the punched cards themselves cannot predict. The efficiency of this approach depends upon the adequacy of the coding schemes, the completeness of the coverage, the accuracy of the encoding and decoding of information, and skill in framing questions and processing answers.

The Coordination Center has now been in existence for six years, supported almost entirely by public funds.[1] Its development has been fostered by the counsel of more than sixty scientists, expertly representing several disciplines, who generously contributed their time to committee work. During this preliminary period the classifications for coding information were prepared and tested and mechanisms for processing data were initiated. Other techniques for expediting the correlation of chemical structure and biological activity have been explored. Some of these have been abandoned as impracticable or as diversions from the main activity, but others are being continued.

The Coordination Center suffers from the difficulties attending any pioneering venture. There is no unanimity of opinion as to what constitutes adequate coding or, in other words, what the unit of thought should include. The chemical unit of thought, being descriptive, is relatively simple. Molecular formulas and structural groupings readily lend themselves to classification. If it becomes necessary to catalogue spatial relationships or types of chemical activity, as is very likely, the task will be augmented. Biological activity, being dynamic, does not readily lend itself to classification. Organisms, organs, tissues, and cells can be catalogued reasonably well. Fundamental actions, however, are subject to some interpretation, and different disciplines use different terminology for the same concept or the same word for different activities.

A code, therefore, must express certain compromises in terminology, standardized by rules and definitions to assure uniformity in coding. In studies on mode of chemical action the most specific response is important, but in studies on therapeutants or practical pesticides the syndrome or gross effect may be of chief interest. There is no general agreement as to what constitutes an "effective" compound, and this calls for quantitative expressions that identify in some way the degree of activity. Biological variation, aggravated by different conditions of test, necessitates the expression of certain of these conditions. The code that has been developed takes these various items into account and serves somewhat as a guide to the coder as to what data to encode. Biological data are expressed in so many different ways, however, that the coder must exercise some judgment. This means that people with specialized training must do the coding, and the coder must be methodical in temperament and sympathetic with the goals to be achieved.

Complete coverage of the literature and coding of available unpublished data are a staggering task—impossible to achieve in the foreseeable future with the budget in sight. It is believed, however, that partial coverage, suggested by the random sample technique, can be very useful and in any case can serve to determine the possible utility of such a research tool. On this basis the center is proceeding, selectively sampling the data in such a way as to favor those areas in which service is likely to be most productive.

Now the center is entering a new phase—one of expanding its files, improving its efficiency, and evaluating its potential scientific usefulness. Although predicated on long-range vision, its program cannot remain visionary. Without efficiency or demonstration of practical values, the center will not attract funds that will ensure its ultimate success. Without meeting a real scientific need, its existence will not be justified. Upon viewing the center's limited experience in answering questions one can find examples demonstrating that its techniques provide an entirely new type of research tool of great potential value. Other examples demonstrate the possibility of economy of scientific effort. Whether these potentialities and possibilities will be realized depends upon the center's efficiency, on the one hand, and its acceptance as a useful instrument by the scientific fraternity, on the other. In any case it is a significant adventure in documentation—an attempt to do something about a much-discussed problem. The center might well serve as an experimental type for many other areas of science where the problem is similar and just as urgent.

[1] Supplied by the Army, Navy, Public Health Service, Atomic Energy Commission, and the American Cancer Society.